Supplement for Where to Add Actions in Human-in-the-Loop Reinforcement Learning

Appendix

Details of Optimistic Estimation Methods

UCRL This strongly optimistic approach is based on UCRL2 (Jaksch, Ortner, and Auer 2010), which defines a confidence set over each transition/reward distribution, and takes the maximum valid distribution in the confidence set when planning. Specifically, in a finite-horizon setting it allows the L1 norm of the transition distribution to deviate from the MLE by at most $\sqrt{\frac{14\log(SA_\ell\tau\ell/\delta)}{\max(N,1)}}$, where ℓ is the number of episodes, N is the number of transition samples, and δ is a user-specified confidence parameter. UCRL's bound incorporates global uncertainty to ensure that the true MDP is within the confidence set with high probability (Auer and Ortner 2007). We use this bound to quantify the uncertainty over our outcome distribution, setting $\delta = 0.05$ as in Osband et al. 2013.[1]. The advantage of UCRL2's constraint on the L1 is that it is easy to calculate the most optimistic distribution that obeys the constraint, as explained in Strehl and Littman 2004. Since we are in a finite horizon setting, we calculate the most optimistic distribution for each (s, t) pair.

MBIE Model-based Interval Estimation (MBIE) (Strehl and Littman 2004; 2008) is a very similar idea to UCRL, but simply bounds the local L1 divergence at each state with probability $1-\delta$. The original MBIE algorithm (Strehl and Littman 2004) used a bound on their transition dynamics of $\sqrt{\frac{2(|S|-1)\log(N+1)-\log(\delta)}{N}}$. However, similar to later versions (Strehl and Littman 2008), we use the bound shown by Weissman et al. 2003, namely that for a discrete distribution with O outcomes, the probability of the error in L1 divergence being ϵ or greater after N samples is at most $(2^O-2)e^{-N\epsilon^2/2}$, to derive the refined bound of $\sqrt{\frac{2\log((2^O-2)/\delta)}{N}}$. The advantage of this bound is that it decreases faster with the number of samples N and matches our outcomes setting. Again we let $\delta = 0.05$ corresponding to 95% confidence, and use $\max(N, 1)$ to deal with the 0-sample case.

(Optimistic) Thompson Sampling Thompson sampling (Thompson 1933), also known as posterior sampling, has been shown to be one of the best performing algorithms empirically at handling the exploration/exploitation tradeoff (Chapelle and Li 2011; Osband, Russo, and Van Roy 2013). Unfortunately, it is not optimistic. Although this tends to be a strength in a explore/exploit setting where it tends to boost exploitation, here it is a major weakness as if the sample happens to be pessimistic we may wastefully add an action at a state where an existing action may already be good. One previously proposed way of partially alleviating this problem is Optimistic Bayesian Sampling, (May et al. 2012) which rejects samples with values lower than the mean of the distribution.[2] The mean of a Dirichlet with parameters $\alpha_1, \ldots \alpha_O$ is a vector X with components $\frac{\alpha_i}{\sum_j \alpha_j}$, so we reject sampled distributions with values of less than $V(X)$. The choice how to set the prior (Dirichlet parameters) is up to the user, however if a Bayesian algorithm such as PSRL (Osband, Russo, and Van Roy 2013) is used, the user will likely have to supply a prior distribution in any case.

BOSS An alternate approach to inject optimism into a posterior sample is simply to sample J times and take the sample with maximum value. This is the key idea behind the Best of Sampled Set (BOSS) algorithm (Asmuth et al. 2009), which showed this approach can have attractive theoretical properties in a traditional explore/exploit setting. Unfortunately, a significant downside of BOSS is that it is unclear how to set J. Asmuth et al. 2009 found a value of $J = 10$ to work best empirically so we use that value without further tuning.

[1] Note that since it is not immediately clear how to translate the UCRL analysis to an MDP specified in terms of outcomes, we simply use apply UCRL's bound on the transition distribution to our outcome distribution. Also, even though it seems as though $\delta = 1/\ell$ is needed to induce sublinear expected regret for UCRL, we follow Osband et al. 2013 for the parameter values.

[2] OBS as proposed by May et al. 2012 takes the max of the sample and the mean, we instead resample if the current sample is lower than the mean to further increase optimism. Additionally, since we have Dirichlet distributions we must be careful in distinguishing the mean of the posterior distribution with the expected final value. However, since the values of each outcome are considered to be a constant and not dependent on the outcome distribution, it suffices to dot product the mean of the Dirichlet with the vector of values.

Proof of Lemma 1

Lemma 1. (Non-starving) *Consider ELI using a prior distribution f which consists of an independent Dirichlet prior on outcome distributions of $\alpha_i = c$ for some $c > 0$. Assume for a given $\epsilon > 0$, after N_ϵ actions are added to each state, additional actions improve the value of each state by at most ϵ. Let $\mathcal{C}$ be an arbitrary class of models with N_ϵ actions which has non-zero probability under our chosen prior. Assume that the true model M for the first N_ϵ actions is drawn from $\mathcal{C}$ according to $f(M|\mathcal{C})$.[3] Finally, assume[4] that for each s there exists $o_1, o_2 \in \mathcal{O}$ such that $T(s, o_1) = T(s, o_2); R(s, o_1) \neq R(s, o_2)$. Then, as the number of actions added by ELI goes to infinity, our ELI approach will eventually uncover actions at each state such that the optimal policy in the MDP with added actions is at least ϵ-optimal (with respect to the full set of actions).*

Proof. For a contradiction, assume there is some non-vanishing probability that the optimal policy in the MDP with added action is less than ϵ-optimal. For this to be the case, clearly there must be a state s and timestep t such that $|A_{s,\ell}| < N_\epsilon$ as $\ell \to \infty$ and $V(s,t)$ is not epsilon-optimal, but if we added additional actions to that s, $V(s,t)$ in the MDP with added actions would be epsilon-optimal. Since the number of total actions added by ELI goes to infinity, there must be at least one other state $s' \neq s$ such that the number of added actions added by ELI at s' go to infinity. Since our ELI method selects the state with the highest ELI score, there must be an infinite number of rounds where the ELI score of s' is greater than that of s. Now, since the ELI score for s' contains a factor of $\frac{1}{|\mathcal{A}_{s',\ell}|+2}$ the score will go to zero as the number of actions at state s' goes to infinity. Therefore, the only way for us to add a finite number of actions to s is for the ELI score of state s to go to zero as well. Since we sum over timesteps and all ELI scores are nonnegative, this means the score of s at time t must go to zero as well.

Now, we know the ELI score for state s at time t is $\frac{1}{|\mathcal{A}_{s,\ell}|+2}(V_{max}(s,t) - \hat{V}(s,t|\mathcal{A}_{s,\ell}))$, or, since we know $|A_{s,\ell}| < N_\epsilon$, at least $\frac{1}{N_\epsilon+2}(V_{max}(s,t) - \hat{V}(s,t|\mathcal{A}_{s,\ell}))$. Clearly, the only way for this to go to zero is for $\hat{V}(s,t|\mathcal{A}_{s,\ell})$ to go to $V_{max}(s,t)$, which means for some $a_s \in \mathcal{A}_{s,\ell}$ $\hat{Q}(s,a_s,t)$ converges to $V_{max}(s,t)$.

By assumption, there exists $o_1, o_2 \in \mathcal{O}$ such that $T(s,o_1) = T(s,o_2); R(s,o_1) \neq R(s,o_2)$. Without loss of generality, label o_1 and o_2 such that $R(s,o_1) > R(s,o_2)$. Clearly, in order for $\hat{Q}(s,a_s,t)$ to approach $V_{max}(s,t)$ the estimated $\hat{P}(o_2|s,a_s)$ must go to zero as ℓ increases, or we could have increased $\hat{Q}(s,a_s,t)$ by shifting the probability mass to $\hat{P}(o_1|s,a_s)$. If we sample (s,a_s) a finite number of times, the BOSS sampling procedure will sample $\hat{P}(o_2|s,a_s) > 0$ with some positive and non-vanishing probability. If we sample $\hat{P}(o_2|s,a_s)$ an infinite number of times, it will converge to its true value. The probability of any outcome distribution placing zero probability on any outcome under the posterior MDP distribution $f(M)$ is zero, and thus likewise under $f(M|\mathcal{C})$. Since the distribution over outcomes at (s,a_s) was sampled according to $f(M|\mathcal{C})$, $P(o_2|s,a_s) > 0$ with probability 1 in the true (sampled) model. So $\hat{Q}(s,a_s,t)$ does not converge to $V_{max}(s,t)$. □

[3]This assumption on the relationship of the true model to the prior is fairly weak, and similar to that used in the analysis of PSRL (Osband, Russo, and Van Roy 2013).

[4]This assumption is due to the considered outcome setting, in fact in discrete MDPs Osband et al. 2013 proposes priors which treat rewards and transitions independently which would imply a stronger assumption.

Simulation Domains

Riverswim (Strehl and Littman 2008) is a chain MDP with 6 states and 2 ground actions per state that requires efficient exploration (Osband, Russo, and Van Roy 2013). For a diagram and complete description of the environment, see Osband et al. 2013. Similar to past work (Mandel et al. 2016) we used a horizon of 20 and use 5 relative outcomes for moving left and right or staying with some reward. We used a flat ($\alpha_i = 1$) Dirichlet prior over outcome distributions for PSRL.

Marblemaze (Asmuth et al. 2009; Russell and Norvig 1994) is a gridworld MDP with 36-states and 4 ground actions per state that allows a significant amount of prior knowledge to be encoded in the outcomes framework (Asmuth et al. 2009; Mandel et al. 2016). For a diagram and complete description of the environment, see Asmuth et al. 2009. As in past work (Mandel et al. 2016), we used a horizon of 30 and a set of 5 outcomes denoting whether the agent moved in each cardinal direction or hit a wall (in keeping with past work, the coordinates of the goal and pits are assumed to be known). We also set the reward for falling in a pit to be -0.03. As in riverswim, we used a flat ($\alpha_i = 1$) Dirichlet prior over outcome distributions for PSRL.

We also try a variant of Marblemaze with an uninformative outcome space. Here, the outcome space consists of every possible combination of the 3 possible reward values and the 31 possible valid next states (plus the terminal signal for falling into one of the four pits or reaching the goal). To allow PSRL to learn faster in this large outcome space, we used a Dirichlet prior of ($\alpha_i = \frac{1}{O}$) over outcome distributions, which encourages sparsity.

The Large Action Task was introduced by Sallans et al. 2004 as a testbed for algorithms that cope with an action space too large to explore directly. In this setting, states and actions are described by vectors of K bits.[5] On each run the environment is initialized by picking 13 bit vectors uniformly at random to represent the states of the problem. Then each of these states is associated with another bit vector (again chosen uniformly at random) to represent the optimal action. At the start of an episode a fresh bit vector is selected uniformly at random, and the the agent starts the episode at the closest (in hamming distance) state to that vector.[6] The agent picks a bit vector as an action and re-

[5]Sallans et al. considered cases where the number of bits for states and actions differed, but we keep them the same for simplicity.

[6]For implementational reasons we do not actually sample a

ceives a reward equal to the number of bits that matched between the chosen action and the optimal action minus $K/2$.[7] In Sallans et al. the environment was closer to a contextual bandit setup, where after each timestep a new key state was drawn independent of the action taken. To make the problem more interesting from a reinforcement learning perspective, we deterministically transition by XORing the chosen action and current state vector and finding the closest (in hamming distance) next key state. In our experiments we choose $K = 20$, so there are only 13 states but 2^{20} ground actions. We use a horizon of 5, and the outcome space is uninformative with one outcome for each key state and reward, so 273 total outcomes. Due to the large outcome space we use a sparse prior, similar to large outcomes version of marble-maze ($\alpha_i = \frac{1}{O}$).

Improving experts in the large action task behave as follows. The initial action for s is drawn by first uniformly sampling a hamming distance, and then uniformly sampling a vector with that hamming distance from the optimal action for s. Subsequent actions are generating by either (with 50% probability) generating another random vector with the same hamming distance as the closest currently available action for s, or (with the remaining 50% probability) choosing an improved hamming distance uniformly at random and then choosing a vector with that hamming distance at random.

Poor experts in the large action task behave as follows. The initial action is generated uniformly at random, and subsequent actions are generated with probability proportional to their hamming distance to the optimal action, so poor actions are much more likely to be generated.

References

Asmuth, J.; Li, L.; Littman, M. L.; Nouri, A.; and Wingate, D. 2009. A Bayesian sampling approach to exploration in reinforcement learning. In *UAI*, 19–26. AUAI Press.

Auer, P., and Ortner, R. 2007. Logarithmic online regret bounds for undiscounted reinforcement learning. *NIPS* 19:49.

Chapelle, O., and Li, L. 2011. An empirical evaluation of Thompson sampling. In *NIPS*, 2249–2257.

Jaksch, T.; Ortner, R.; and Auer, P. 2010. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research* 11(Apr):1563–1600.

Mandel, T.; Liu, Y.-E.; Brunskill, E.; and Popović, Z. 2016. Efficient bayesian clustering for reinforcement learning. In *IJCAI*. AAAI Press.

May, B. C.; Korda, N.; Lee, A.; and Leslie, D. S. 2012. Optimistic bayesian sampling in contextual-bandit problems. *Journal of Machine Learning Research* 13(Jun):2069–2106.

fresh vector every time, instead we sample 100 vectors when the environment is created to construct an approximate distribution over key states and sample from that at the start of each episode.

[7] Sallans et al. did not subtract off $K/2$ but we do so to get a better sense of the difference among algorithms in terms of cumulative reward, as this linear transform does not meaningfully change the problem but simply normalizes rewards so that random achieves zero reward.

Osband, I.; Russo, D.; and Van Roy, B. 2013. (More) efficient reinforcement learning via posterior sampling. In *NIPS*, 3003–3011.

Russell, S., and Norvig, P. 1994. *Artificial Intelligence A Modern Approach*. Englewood Cliffs, NJ: Prentice Hall.

Sallans, B., and Hinton, G. E. 2004. Reinforcement learning with factored states and actions. *Journal of Machine Learning Research* 5(Aug):1063–1088.

Strehl, A. L., and Littman, M. L. 2004. An empirical evaluation of interval estimation for Markov decision processes. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, 128–135. IEEE.

Strehl, A. L., and Littman, M. L. 2008. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences* 74(8):1309–1331.

Thompson, W. R. 1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 285–294.

Weissman, T.; Ordentlich, E.; Seroussi, G.; Verdu, S.; and Weinberger, M. J. 2003. Inequalities for the L1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*.